冯广, 汤翀. 基于语义融合特征的多人像语义分割方法[J]. 广东工业大学学报, 2025, 42(2): 20–28. doi: 10.12052/gdutxb.230211.
Feng Guang, Tang Chong. A multi-portrait semantic segmentation method based on semantic fusion features[J]. Journal of Guangdong University of Technology, 2025, 42(2): 20–28. doi: 10.12052/gdutxb.230211.

基于语义融合特征的多人像语义分割方法

冯　广[1]，汤　翀[2]

（1. 广东工业大学　自动化学院，广东　广州 510006；2. 广东工业大学　计算机学院，广东　广州 510006）

摘要： 人像语义分割是计算机视觉领域的重要研究内容之一，但现有的人像语义分割方法容易忽略多人像图像中的小尺寸人像。同时分割结果中容易出现多个人像之间相互粘连的现象。再者，图像中人像之间存在相互遮挡现象容易导致人像边缘分割精度不佳。基于以上问题，本文提出一种融合标签语义的多人像语义分割方法，对图像中的多个人像分配多个标签，并将语义标签嵌入同时作为编码器的输入，使用跨模态交叉注意力模块对语义标签和图像特征表示进行相关性建模，将语义融合的特征表示作为模型每一层编码器的输出。提出HRF attention模块，基于目标检测算法对图像生成的多个假设分别进行特征提取。将该网络在 Supervisely增强数据集上训练测试，实验结果表明该算法模型在3个评估指标PA、MIoU、Dice上分别达到95.94%、94.60%、96.02%的精度，较语义分割模型U-net、PSPNet、Deeplab v3+、PortraitNet、Swin Unet具有更高的分割精度。

关键词： 语义分割；U-net；多标签；注意力机制；跨模态特征融合

中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1007–7162(2025)02–0020–09

A Multi-Portrait Semantic Segmentation Method Based on Semantic Fusion Features

Feng Guang[1], Tang Chong[2]

(1. School of Automation, Guangdong University of Technology, Guangzhou 510006, China; 2. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Portrait semantic segmentation is one of the important research contents in the field of computer vision, but the existing portrait semantic segmentation methods are liable to ignore the small size portraits in multi-person portrait images. At the same time, the segmentation results are prone to the phenomenon of mutual adhesion between multiple portraits. Moreover, the phenomenon of mutual occlusion between portraits in the image easily leads to poor segmentation accuracy of portrait edges. Based on the above problems, a semantic segmentation method for multiple portraits with fused label semantics is propose, where multiple labels are assigned to multiple portraits in an image, and semantic labels are embedded as inputs to the encoder at the same time, and the semantic labels and the image feature representations are correlated using the cross-modal cross-attention module, and the semantically fused feature representations are obtained as outputs of the encoder at each layer of the model. The HRF attention module is proposed to generate multiple hypotheses for image based on target detection algorithm for feature extraction separately. The network is trained and tested on Supervisely augmented dataset. The experimental results show that the algorithmic model achieves 95.94%, 94.60%, and 96.02% accuracy on the three evaluation metrics of PA, MIoU, and Dice, respectively, and has higher segmentation accuracy than the semantic segmentation models U-net, PSPNet, Deeplab v3+, PortraitNet, and Swin Unet.

Key words: semantic segmentation; U-net; multi-label; attention mechanism; cross-modal feature fusion

语义分割是计算机视觉领域的主要研究内容之一，人像语义分割则是语义分割的重要组成部分，其目的是在含有人像的图片中将人像语义区域和复杂的背景信息语义区域分割开，获得一个良好的人像

收稿日期： 2023-12-27　　**录用日期：** 2024-07-17
基金项目： 国家自然科学基金资助项目(62237001)
作者简介： 冯广(1973–)，男，教授，博士，主要研究方向为深度学习、网路控制，E-mail：von@gdut.edu.cn
通信作者： 汤翀(1998–)，男，硕士研究生，主要研究方向为图像分割，E-mail：1412562728@qq.com

边界[1]，从而进行后续的处理。深度学习领域的卷积神经网络(Convolutional Neural Networks, CNN)可以从训练图像中自动学习不同混合的表征特征，从而对图像进行语义分割[2]。其中以U-net[3]为代表的基于编码器解码器的CNN语义分割模型在较小的数据集上就能表现出较好的效果，是目前语义分割领域的主流选择，但现有的基于U-net的语义分割方法在面对分割对象严重未对齐或被遮挡的图像时相对不准确。针对此类问题，Wei等[4]提出HCP(Hypotheses CNN Pooling)，先通过对象检测技术生成一组候选对象窗口，再将其输入至模型中进行特征提取，从而将多标签问题转换为多个单标签任务。但此类方法可能会造成特征表示中图像全局空间信息的丢失，导致同一张图像中的不同尺寸人像目标边界分割效果不佳。针对此类问题，Reza等[5]在编码器引入空洞空间金字塔池化模块(Atrous Spatial Pyramid Pooling，ASSP)，通过并行多个分支异扩张率的膨胀卷积，来抽取不同大小感受野的多尺度特征。在这基础上，赵为平等[6]通过将ASSP模块中的全局平均池化改进为混合条带池化，进一步提高模型学习多尺度信息的能力，更好地关注图像中不同尺寸目标对象细致的图像边界。同时，学界还通过使用注意力机制解决此类问题。其中Yang等[7]提出在U-net每一个编码器的后面连接一个残差注意力模块，通过显式建模卷积特征通道之间的相互依赖性来提高特征表示能力。在此基础上，还有学者通过对多个注意力机制进行连接，捕获编码器输出特征表示中不同维度的空间信息，进一步提高分割结果的准确度[8-9]。尽管学界已经就U-net进行了多种改进，但上述的人像语义分割方法都使用单标签配置，即为多个人像赋予同一个人像标签。然而，现实中的人像图像往往包含多个人像、多标签图像中人像之间的不同组成和交互，例如部分可见性和遮挡，使得这需要更多的注释数据来覆盖不同的情况[10]，导致目前单标签配置下的语义分割模型分割精度不佳，需要对网络结构继续进行改进和优化。

对此，本文提出一种融合标签语义的人像语义分割算法，主要包括以下工作：(1) 采取多标签分割的模式，为数据集中的多人像图像的多个人像分配多个标签，并将语义标签嵌入作为编码器输入，与图像的特征表示进行特征融合，获取带有语义联合的特征表示。(2) 提出HCF attention (Hypotheses CNN Fusion attention)模块对图像进行特征提取，使用目标检测算法模型生成多个假设，再将多个假设作为输入，进行特征提取并进行特征融合。(3) 使用跨模态交叉注意力机制进行特征融合，将语义标签嵌入和图像特征表示分别进行自注意力操作并进行信息交互，获取联合特征，再使用该联合特征调整两个模态下的特征表示权重进行特征融合。

1　网络框架

1.1　网络总体框架

本文所使用的模型在U-net编码器−解码器结构的基础上进行设计，主要分为3个部分：编码器(Encoder)、跳跃连接(Skip connection)、解码器(Decoder)，如图1所示。

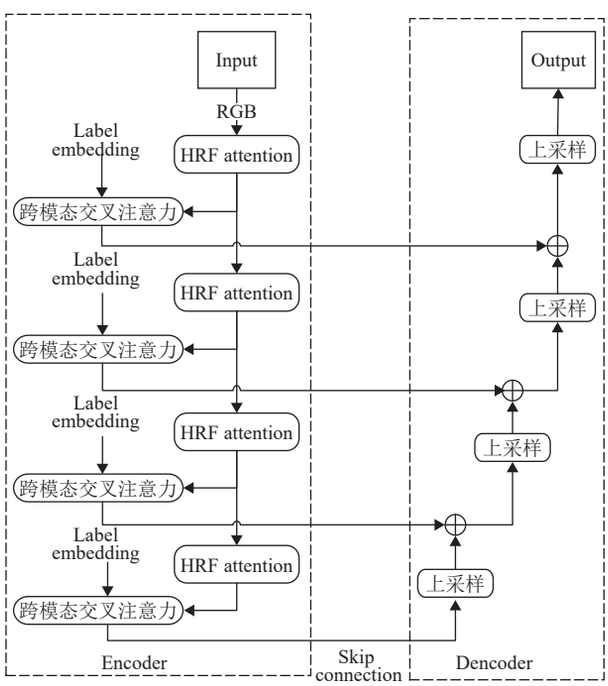


图 1　网络结构

Fig.1　Network architecture

编码器阶段主要分为2个输入，RGB图像输入及语义标签嵌入。RGB图像输入分支中，使用假设残差融合注意力(Hypotheses Residual Fusion Attention，HRF Attention)模块获得对应的图像特征表示。然后将语义标签嵌入和图像特征表示输入至跨模态交叉注意力模块中，进行二者的相关性建模，以获得语义联合的特征表示，作为该层编码器的输出。经过4个连续的编码器块后，获得输入图像的深层特征，再输入至解码器中进行上采样操作。跳跃连接阶段，保持原U-net中结构，把每一层编码器的输出和解码器输出进行连接，使深层和浅层的信息融合起来。

解码器阶段由4个连续的上采样操作组成，每一个上采样操作都使用3×3逆卷积的方式，将图像恢复

至相应的尺寸。同时，本文将U-Net结构的特征图级联操作改为逐像素相加，即每一层编码器的输出与上一层上采样的结果进行逐像素相加后再进行上采样操作。通过此类方法可以有效减小模型维度[12]。

1.2 HRF Attention模块

本文提出HRF Attention模块作为图像输入的特征提取模块，如图2所示。其主要分为使用目标检测算法生成多个假设，并对假设进行特征提取和特征融合的HRF(Hypotheses Residual Fusion)分支，以及针对整张图像进行特征提取，并使用混合注意力进行特征加权的注意力分支，其结构如图2(a)所示。

在HRF分支中，首先将图像输入至目标检测算法模型中，获取对应的候选窗口，生成一定数量的假设(Hypotheses directly)。本文选用YOLO V5预训练模型[13]执行目标检测算法，选定目标并生成假设。其过程如图3所示，其中图3(a)列为输入图像，图3(b)列为生成候选框，图3(c)列为根据候选框生成的假设。

再将生成的多个假设输入至空间残差块中进行特征提取，此类方法可以将问题转变为多个单标签任务，可以更好地发挥 CNN 模型的强大判别能力。空洞残差块(Atrous Residual Block)的结构如图2(b)所示。不同于常规残差块，空间残差块使用3个并行的不同扩张率(d_1=1，d_2=2，d_3=3，d为扩张率)的空洞卷积替代常规残差块中卷积核大小相同的连续卷积。在空洞残差块中，会首先聚合具有高膨胀率的特征表示，再聚合具有较小膨胀率的特征表示，从而更好地保留多尺度上下文中的层次依赖性[14]。同时，并行的空洞卷积可以抽取不同大小感受野的多尺度特征，能较好地关注图像中目标对象细致的图像边界。残差分支则可以避免梯度消失的问题[15]。得到每一个假设对应的特征表示后，再将其输入至特征融合模块(Feature fusion)中，结构如图2(c)所示。特征融合模块首先会将每一个输入的假设特征表示(Hypotheses feature)进行全局平均池化(Global Average Pooling, GAP)，再通过Attention操作，获得对应的注意力权重。然后，将该注意力权重和原本的输入特征表示进行相乘，得到每一个假设调整权重的特征表示，最后进行特征拼接。Attention操作通过分配权重的形式，有效降低非人像目标假设的权重，避免一些无用的假设的影响。最后，使用1×1卷积对融合后的图像特征映射进行降维，作为HRF分支的输出。

在注意力分支中，首先将输入图像输入至3×3卷积中，获取整张图像的初始特征表示。再经过混合注意力对整张图像的特征表示赋予更多的全局空间信息。本文使用的混合注意力如图4所示，其中上半部

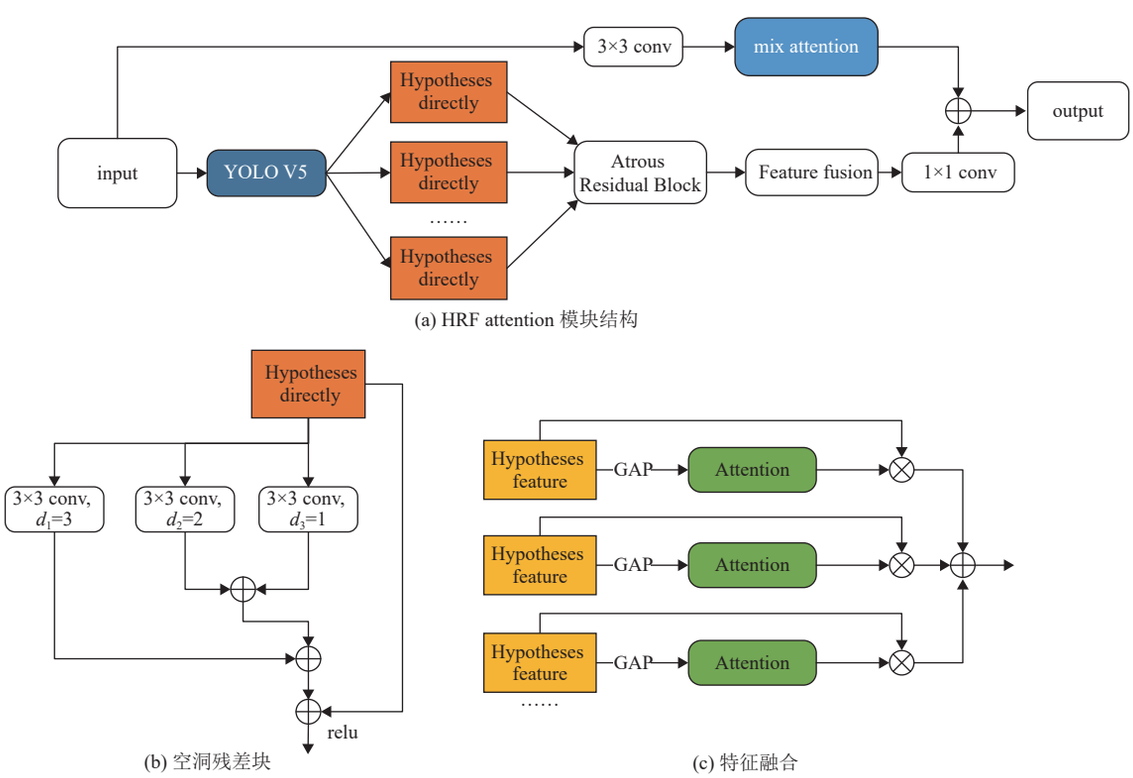


图 2 图像特征提取模块结构

Fig.2 Structure of image feature extraction module

图 3　假设生成过程

Fig.3　Hypothesis generation process

分是通道注意力模块，下半部分是位置注意力模块。通道注意力机制通过学习不同通道的权重，可以降低无效通道的响应，从而抑制无用的特征激活；而位置注意力机制通过对任意两个像素点进行建模来表示两个位置的相关性，这样具有相似特征的两个像素可以相互提升，得到较高的权重。

在通道注意力模块中，输入特征$\boldsymbol{A}$首先经过1×1卷积进行降维操作，再使用平均池化(Avgpool)和最大池化(Maxpool)来聚合输入的空间信息，并经过多层感知机MLP，得到两个空间上下文特征$\boldsymbol{B}$、$\boldsymbol{C}$。然后，将两个空间上下文特征进行合并，计算得到相应的通道注意力图$\boldsymbol{M}$，如式(1)所示。

$$\boldsymbol{M}=\sigma(\mathrm{MLP}(\boldsymbol{\mu})+\mathrm{MLP}(\boldsymbol{\nu})) \tag{1}$$

式中：σ为sigmoid函数，多层感知机MLP隐藏激活大小设置为 $R^{C/r\times1\times1}$，C为通道数，r为缩减比率[16]，$\boldsymbol{\mu}$和$\boldsymbol{\nu}$分别为经过AvgPool和MaxPool后得到的向量。最后，将该注意力图$\boldsymbol{M}$和输入特征进行相乘，得到通道注意力特征。

在位置注意力模块中，对于输入特征，首先经过1×1卷积进行降维操作，得到与输入特征尺寸相同的特征$\boldsymbol{D}$、$\boldsymbol{E}$。再对特征$\boldsymbol{D}$进行转置后和特征$\boldsymbol{E}$进行相乘，进而建模特征图中任意两个像素点之间的相似性，计算得到相应的位置注意力图$\boldsymbol{P}_{ij}$，其中计算如式(2)所示。

$$\boldsymbol{P}_{ij}=\frac{\exp(D_i\times E_j)}{\sum_{i=1}^{N}\exp(\underset{i}{D}\times\underset{j}{E})} \tag{2}$$

式中：$N=H\times W$，H和W分别为输入特征表示的高和宽，j和i分别为特征图中的第j个像素和第i个点。然后，将该注意力图$\boldsymbol{P}_{ij}$和输入特征进行相乘，得到位置注意力特征$\boldsymbol{P}$。如式(3)所示。

$$\boldsymbol{P}=\delta\sum_{j=1}^{N}\sum_{i=1}^{N}\boldsymbol{P}_{ij}\times\boldsymbol{A}_j \tag{3}$$

式中：δ为尺度系数，初始化为0，通过模型不断迭代逐渐获得对应的权重。最后，将该注意力图$\boldsymbol{P}$和输入特征进行相乘，得到坐标注意力特征。

对两个注意力输出特征分别使用1×1卷积还原至输入维度，最后进行特征融合，得到混合注意力模块输出特征。

最后，将经过混合注意力模块后的整张图像的特征表示和各假设进行特征融合后的特征表示进行逐像素相加，获得图像的最终特征表示。

1.3　跨模态交叉注意力模块

语义标签和图像属于不同的模态，不同模态数据对于语义分割的贡献度不同。本节基于自注意力机制，建立了一种跨模态信息交互模块，可以对语义标签和图像特征表示两种模态之间的相关性进行建

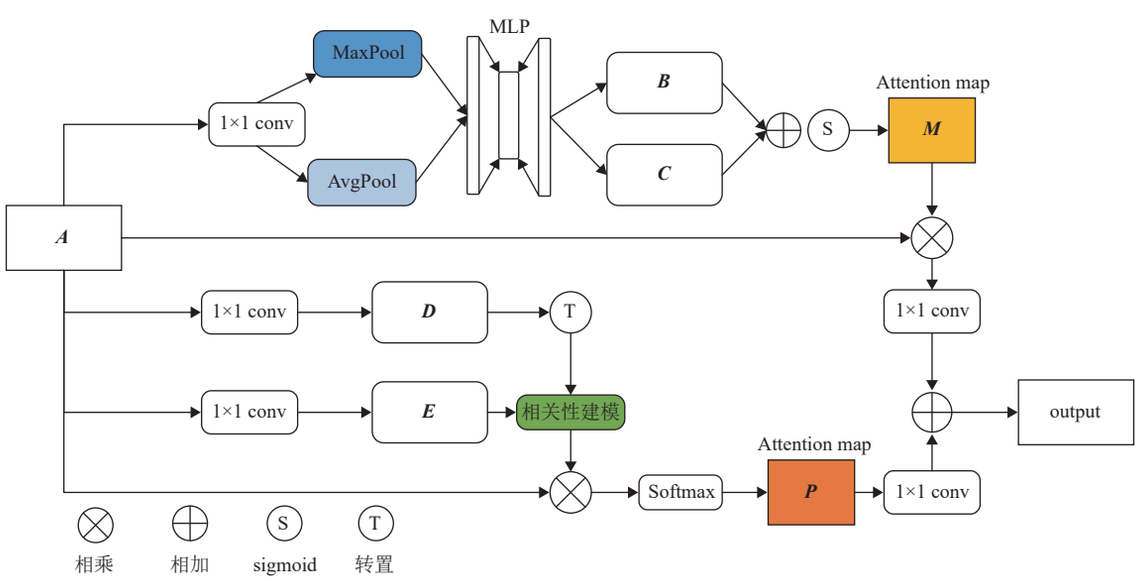
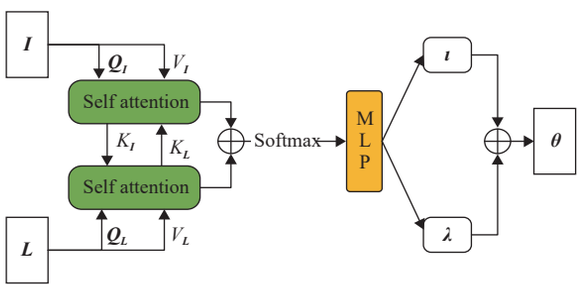


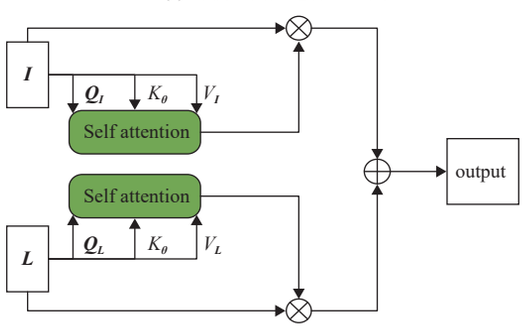
图 4　混合注意力结构

Fig.4　Structure of mixed attention

模[17]，获得两种模态下的联合特征表示。不同于常规的跨模态融合模块，跨模态交叉注意力模块会进行两次特征融合。在获得经过第一次信息交互的联合特征表示后，不会将联合特征作为最终的输出，而是再通过联合特征表示，对两种模态的特征进行指导，再进行二次特征融合，获取最终的语义特定的图像特征表示。其结构如图5所示。



(a) 跨模态向量进行信息交互



(b) 基于联合特征进行信息指导

图 5 跨模态交叉注意力模块

Fig.5 Cross-modal cross-attention block

在跨模态交叉注意力机制中，对语义标签嵌入和图像特征表示进行相关性建模，获得联合特征$\boldsymbol{\theta}$，其过程如图5(a)所示。首先，对输入向量$\boldsymbol{I}$和$\boldsymbol{L}$进行交叉自注意力操作，获取对应的注意力向量$\boldsymbol{F}$，此处采用式(4)计算对应特征向量的自注意力向量$\boldsymbol{F}_I$及$\boldsymbol{F}_L$。

$$\begin{aligned}\boldsymbol{F}_I &= \mathrm{Softmax}\left(\frac{K_L \boldsymbol{Q}_I^{\mathrm{T}}}{\sqrt{d}}\right) V_I \\ \boldsymbol{F}_L &= \mathrm{Softmax}\left(\frac{K_I \boldsymbol{Q}_L^{\mathrm{T}}}{\sqrt{d}}\right) V_L\end{aligned} \tag{4}$$

式中：$\boldsymbol{I}$为图像特征向量，$\boldsymbol{L}$为标签嵌入，$\sqrt{d}$为特征维度，$\boldsymbol{Q}$和$\boldsymbol{V}$分别取自各自输入模态，K取自另一输入模态，计算对应的注意力向量，从而进行输入特征之间的信息交互。

然后，将获得的对应模态下的自注意力向量进行拼接，再通过softmax和多层感知机MLP获得向量$\boldsymbol{\iota}$和$\boldsymbol{\lambda}$，并对向量$\boldsymbol{\iota}$和$\boldsymbol{\lambda}$进行拼接，获得联合特征$\boldsymbol{\theta}$。

尽管联合特征表示是通过两种模态的特征交互所获得，但仍存在信息冗余、丢失以及噪声等问题[18]，为了克服以上问题，本文基于联合特征$\boldsymbol{\theta}$，对语义标签和图像特征两种模态的向量进行指导，过程如图5(b)所示。分别将特征表示$\boldsymbol{I}$和$\boldsymbol{L}$进行联合自注意力操作，其中K和V取自联合特征$\boldsymbol{\theta}$，$\boldsymbol{Q}$分别取自输入向量$\boldsymbol{I}$和$\boldsymbol{L}$，从而计算输入向量$\boldsymbol{I}$和$\boldsymbol{L}$经过联合特征$\boldsymbol{\theta}$指导后的注意力权重。然后，再与本模态下的输入进行相乘，获得权重调整后的特征表示。再将两个模态下新的特征表示进行拼接，获得最终的语义特征的的图像特征表示。

1.4 混合损失函数

Bischk等[19]提出了多任务损失方法来增强最终的语义分割结果，并利用不同的数据函数对数据结果进行不同方面的约束，达到了约束模型输出的良好效果。受此启发，为了使得本文模型在边缘分割精度上会有所提升，本文通过基于像素损失函数和结构相似性损失函数两个损失函数组成混合损失函数来优化语义分割结果，这意味着模型的输出受到两个损失函数的约束。其中，基于像素损失函数主要用于优化像素损失；结构相似性损失函数主要用于优化边界损失。本文所使用的损失函数$\boldsymbol{L}$构成如式(5)所示。

$$\boldsymbol{L} = \alpha \boldsymbol{l}_{\mathrm{pb}} + \beta \boldsymbol{l}_{\mathrm{ssim}} \tag{5}$$

式中：$\boldsymbol{l}_{\mathrm{pb}}$为基于像素损失函数(pixel-based loss)，$\boldsymbol{l}_{\mathrm{ssim}}$为结构相似性损失函数(structural similarity, ssim)。α 和 β 为偏置参数($\alpha+\beta=1$)，主要目的是逐步提取信息并对分割的边缘信息和总精度信息进行初步评估和约束。

像素损失函数主要用于比较两幅图之间像素级的差别，具体可以分为两部分：第1部分是像素携带的信息总量，用二值交叉熵表示；第2部分基于像素值差距测量，成对比较分割结果和人工生成的标签之间的差异。其定义为

$$\boldsymbol{l}_{\mathrm{pb}} = 1 - \frac{1}{\binom{N}{2}} \sum_{i,j,i\neq j} \left[\boldsymbol{I}\left(l_i^S = l_j^{S_v}\right) p_{ij} + \boldsymbol{I}\left(l_i^S \neq l_j^{S_v}\right) p_{ij}\right] \tag{6}$$

式中：N为像素数量，S_v为分割标签的集合，S为基准真实值的集合，p_{ij}为概率。根据经验，p_{ij}根据所有划分的像素对的平均值计算而得，取值在0和1之间，其中0表示分割结果与分割标签相反，1表示分割结果中每个像素都被准确分类。

结构相似性损失函数主要用于衡量预测样本与真实样本之间相似程度的指标。由于其考虑了每个

像素的局部邻域，可以将较高的权重分配给边界。其定义为

$$\boldsymbol{l}_{\text{ssim}} = 1 - \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2\mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{7}$$

式中：x和y分别为两张图片的像素点，μ_x和μ_y为x和y的均值，σ_x和σ_y为x和y的标准差；C_1设定为0.022，C_2设定 为0.042，防止分母为0[20]。

2　实验与结果分析

2.1　实验环境

本文的实验主要在python3.8、pytorch1.7、Cuda 11.0环境下进行编码，在CPU 为Intel i7-11700k、显卡为RTX 3070、内存为16 G的计算机上进行训练。训练选用adam迭代器，设置网络参数初始学习率为 0.01。混合损失函数中的偏置参数α及β经过多次实验后，设定为0.7及0.3(见本文2.4.2节)，batchsize(步长)设为16，迭代训练1 000次。

2.2　数据集

本文数据主要基于公开数据集Supervisely进行搭建。在此基础上，通过对PennFudanPed数据集中的图像进行手工标注掩膜及标签的方式进行数据增强：对于人像个数为2个的图像，根据人像所处位置分配语义标签left person、right person；对于人像个数为3个的图像，根据人像所处位置分配语义标签left person、middle person、right person；对于人像个数多于3个的图像，根据人像所处位置分配语义标签left person、right person。对所有处于中间位置的人像分配语义标签middle person；效果如表1所示。

通过此类数据增强的方式，可以增加数据集中语义标签类型的数量，帮助模型学习到更多的语义标签特征。最终，构建数据集共包含人像图像7 000张，其中重新分配标签图像共1 600张，原始标签图像共5 400张。遵循8:2的划分比例对数据集进行划分，其中5 600张作为训练数据集(包含全部1 600张自标注图像)，1 400张作为测试数据集。

2.3　评估指标

人像语义分割实际上可以转化为人像语义标签区域与背景语义标签区域的分类问题，本文主要使用分割模型常用的像素准确率PA(Pixel Accuracy，像素准确率)、MIoU(intersection over union，平均交并比)和Dice系数评估指标进行评估。计算公式为

$$\text{PA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FT} + \text{FN}} \tag{8}$$

$$\text{MIoU} = \frac{1}{K+1}\sum_{I=1}^{K_v}\frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}} \tag{9}$$

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{10}$$

式(8)和式(9)中：TP为真正例(即测试样本中目标人像能够被准确地预测成目标人像)，FN为假反例(表示测试样本中目标人像的区域能够被准确地预测成人像背景)，FP为假正例(表示测试样本中人像背景能够被预测成目标人像)，TN为真反例(表示测试样本中人像背景能够被准确地预测成人像背景)。式(9)中K为分类的类别(本次实验中为2)。

2.4　实验结果

2.4.1　基于编码器的消融实验

为了验证本文模型中所使用语义标签特征提取

表 1　数据集标注示例

Table 1　Sample of dataset labeling

	原图	掩膜	标签
图像示例1			left person right person
图像示例2			left person middle person right person
图像示例3			left person middle person right person

和图像特征提取模块在编码器中的作用，进行了消融实验，在基准网络模型U-net的基础上设立以下6组网络：Network01代表编码器添加语义标签嵌入，然后将标签嵌入直接和图像特征表示进行融合的模型；Network02代表编码器添加语义标签嵌入，并使用Transformer Encoder对标签嵌入和图像特征表示进行特征融合的模型；Network03代表编码器添加语义标签嵌入，并使用跨模态交叉注意力对标签嵌入和图像特征表示进行特征融合的模型；Network04代表在Network03基础上编码器添加HRF attention模块中的HRF分支，分支使用标准残差块对图像进行特征提取的网络；Network05代表在Network03基础上编码器添加HRF attention模块中的HRF分支，分支使用空洞残差块对图像进行特征提取的网络；Network06代表在Network04基础上编码器添加HRF attention模块中的注意力分支，分支使用混合注意力机制的网络。实验结果如表2所示。

表2 基于编码器的消融实验结果

Table 2 Results of ablation experiments based on encoder

模型	语义标签嵌入	HRF attention 模块，HRF分支(使用标准残差块)	HRF attention 模块，HRF分支(使用空洞残差块)	HRF attention 模块，注意力分支	基于Transformer Encoder进行特征融合	基于跨模态交叉注意力机制进行特征融合	PA/%	MIoU/%	Dice/%
U-net							92.32	90.86	92.78
Network01	√						92.37	90.92	92.83
Network02	√				√		92.56	91.20	93.02
Network03	√					√	92.78	91.37	93.38
Network04	√	√				√	93.75	92.41	94.29
Network05	√		√			√	94.44	93.22	94.85
Network06	√		√	√		√	95.61	94.26	95.88

分析消融实验结果可以看出：与原始模型相比较，编码器引入语义标签嵌入的模型中，Network01、Network02、Network03在3个评估指标PA、MIoU、Dice上都有所有提升。其中使用本文提出跨模态交叉注意力模块对标签嵌入和图像特征表示进行特征融合的Network03，在3个评估指标PA、MIoU、Dice上分别提升了0.46、0.51和0.26个百分点，提升幅度最大。在Network03基础上，继续在编码器添加了HRF attention 模块的Network04、Network05、Network06，在PA和MIoU都呈现较大幅度的提升。其中，在编码器只添加了HRF attention 模块HRF分支的Network04和Network05相较于Network03在PA分别提升了0.97和1.66个百分点，在MIoU分别提升了1.04和1.85个百分点，在Dice分别提升了0.91和1.47个百分点。而使用空洞残差块的Network05对比于使用标准残差块的Network04，在PA、MIoU和Dice上分别有着0.69、0.81和0.56个百分点的优势。而在Network05基础上，在编码器继续添加了注意力分支的Network06相较于Network03，在PA、MIoU和Dice分别提升了2.83、2.89、2.5个百分点，和Network04、Network05相比，具有更大的提升幅度。

2.4.2 基于混合损失函数的超参数实验

为验证本文混合损失函数的有效性，确定偏置参数α及β的最佳取值，进行消融实验。基于U-net基准模型，设置以下6组使用不同损失函数的U-net网络。实验结果如表3所示。

表3 混合损失函数的超参数实验结果

Table 3 Hyper parameter experimental results of mixed loss function

损失函数	α取值	β取值	PA/%
交叉熵损失函数			92.32
混合损失函数	0.1	0.9	92.44
混合损失函数	0.3	0.7	92.51
混合损失函数	0.5	0.5	92.59
混合损失函数	0.7	0.3	92.62
混合损失函数	0.9	0.1	92.58

观察表3可知，比起使用U-net基准损失函数交叉熵损失函数的第1组网络，使用混合损失函数的后5组网络在评估指标PA上都有一定程度提升。而在6组网络中，模型在偏置参数α及β取值0.7和0.3时在PA上取得最优值92.62%。

2.4.3 模型性能对比实验

在本节中，为评估本文所提出算法的性能，将其分别与算法模型U-net[3]、PSPNet[21]、Deeplab v3+[22]、PortraitNet[23]、Swin Unet[24]在评估指标及每张图像平均分割用时上进行对比。实验结果如表4所示。

由表4可以看出，本文提出的算法在PA、MIoU这2个分割指标上表现最优。与U-net基准模型相比，

表 4　不同分割模型性能对比实验结果
Table 4　Comparison of different segmentation models

模型	PA/%	MIoU/%	Dice	平均分割用时/ms
U-net	92.32	90.86	93.78	34.12
PSPNet	93.68	91.95	94.41	40.23
Deeplab v3+	94.46	92.87	95.04	46.55
PortraitNet	94.42	92.80	94.99	32.04
Swin Unet	95.40	94.01	95.82	55.14
本文算法	95.94	94.60	96.02	50.26

PA超出了3.62个百分点，MIoU超出了3.74个百分点，Dice系数超出了2.24个百分点；与基于U-net进行改进的PSPNet和Deeplab v3+相比，PA分别超出了2.26个百分点和1.48个百分点，MIoU分别超出了2.65个百分点和1.73个百分点，Dice系数分别超出了1.61个百分点和0.98个百分点；与轻量型人像分割模型PortraitNet相比，PA提升了1.52个百分点，MIoU提升了1.80个百分点，Dice系数超出了1.03个百分点；与高精度分割模型Swin Unet相比，在3个评估指标PA、MIoU、Dice上也分别有0.54、0.59和0.2个百分点的优势。二者分割精度相差不大，但本文模型与Swin Unet相比，分割速度更快。总体而言，本文研究模型在评估指标上达到了较为可观的精度。

对于一个算法模型而言，模型的可视性结果有利于提升模型的可信度[25]。为了更直观地对本文研究模型的分割性能进行评估，从数据集中选取1组多人像图像进行分割效果可视化对比，各模型可视化分割结果如图6所示。

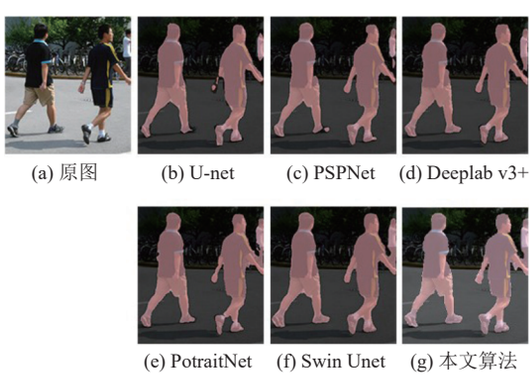


图 6　模型分割可视化对比结果
Fig.6　Model segmentation visual comparison results

从分割结果中观察可知：U-net背景误判较为严重，图6(b)右方人像手臂出现了分割不完整现象；PSPNet基本可以完成多个人像轮廓的分割，但在图6(c)中人像边缘出现了较为明显的锯齿状现象；Deeplab v3+和PortraitNet未出现人像目标分割不完整的现象，但人像与人像之间的边界出现了背景误判现象(如图6(d)中间人像手臂和右方人像脚之间)；Swin Unet分割效果较好，但在面对如图6(f)中间人像对右方小尺寸人像的遮挡现象时，还是会出现人像左脚细节处的分割遗漏现象；本文模型较以上各模型，人像与人像之间的细节分割处理较好，可以较为精细地处理图6(g)中多人像图像中的小尺寸人像，同时，面对图6(g)中相连的右方人像和小尺寸人像，模型也可以对其进行有效分割，人像边缘精度较佳，在小尺寸人像和右方大人像相连处，没有出现误解误判的现象，分割精度符合预期。

3　结论

本文针对多人像图像语义分割中出现的人像之间相互粘连、小尺寸人像目标被忽略、重叠人像边缘细节分割精度不佳等问题，提出了一种融合标签语义的多人像语义分割方法。该方法首先对数据集中的多人像图像进行重新标注，增加数据集中语义标签的数量，然后基于U-net结构进行设计，将标签和图像同时作为编码器输入，获取对应模态下的特征表示后使用跨模态交叉注意力机制进行特征融合，获取语义特定的特征表示作为每一层编码器输出。同时提出HRF attention模块进行图像特征提取，使用目标检测算法生成多个假设作为输入进行特征提取，再进行特征融合，充分发挥CNN的强大判别能力。在Supervisely增强数据集下进行训练测试，本文模型在3个评估指标PA、MIoU、Dice系数分别达到了95.94%、94.60%和96.02%的精度，对比于语义分割模型U-net、PSPNet、Deeplab v3+、PortraitNet、Swin Unet具有更高的精度。但本文模型参数较多，分割速度较其他模型未有较明显优势，后续研究将对模型分割速度与分割精度上的平衡进行更多探索。

参考文献：

[1] 王欣. 基于深度学习的人像分割方法研究[D]. 咸阳: 西北农业科技大学, 2022.

[2] 蒯宇, 王彪, 吴艳兰, 等. 基于多尺度特征感知网络的城市植被无人机遥感分类[J]. 地球信息科学学报, 2022, 24(5): 962-980.
KUAI Y, WANG B, WU Y L, *et al*. Urban vegetation classification based on multi-scale feature perception network for UAV images[J]. Journal of Geo-Information Science, 2022, 24(5): 962-980.

[3] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference. Mu-

nich: Springer International Publishing, 2015: 234-241.

[4] WEI Y, XIA W, LIN M, *et al*. HCP: a flexible CNN framework for multi-label image classification[J]. IEEE, 2015, 38(9): 1901-1907.

[5] AZAD R, ASADI-AGHBOLAGHI M, FATHY M, *et al*. Attention DeepLabv3+: multi-level context attention mechanism for skin lesion segmentation[C] //European Conference on Computer Vision (ECCV) . Glasgow: Springer International Publishing, 2020: 251-266.

[6] 赵为平, 陈雨, 项松, 等. 基于改进的DeepLabv3+图像语义分割算法研究[J]. 系统仿真学报, 2023, 35(11): 2333-2344.
ZHAO W P, CHEN Y, XIANG S, *et al*. Research on image semantic segmentation algorithm based on improved DeepLabv3+[J]. Journal of System Simulation, 2023, 35(11): 2333-2344.

[7] YANG Y, WAN W, HUANG S, *et al*. RADCU-Net: residual attention and dual-supervision cascaded U-Net for retinal blood vessel segmentation[J]. International Journal of Machine Learning and Cybernetics, 2023, 14: 1605-1620.

[8] FU J, LIU J, TIAN H, *et al*. Dual attention network for scene segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146-3154.

[9] AMER A, LAMBROU T, YE X. MDA-unet: a multi-scale dilated attention U-net for medical image segmentation[J]. Applied Sciences, 2022, 12(7): 3676.

[10] 叶庆文, 张秋菊. 采用通道像素注意力的多标签图像识别[J/OL]. 计算机科学与探索. https://link.cnki.net/urlid/11.5602.TP.20230829.1911.004.

[11] 雷俊婷. 基于多尺度特征增强的多标签图像分类研究[D]. 西安: 西北大学, 2023.

[12] 张淑军, 彭中, 李辉. SAU-Net: 基于U-Net和自注意力机制的医学图像分割方法[J]. 电子学报, 2022, 50(10): 2433-2442.
ZHANG S J, PENG Z, LI H. SAU-Net: medical image segmentation method based on U-Net and self-attention[J]. Acta Elect Ronica Sinica, 2022, 50(10): 2433-2442.

[13] WU W, LIU H, LI L, *et al*. Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image[J]. PloS One, 2021, 16(10): e0259283.

[14] CHEN L C, ZHU Y, PAPANDREOU G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[C] //Proceedings of the European Conference on Computer Vision (ECCV) . Munich: Springer International Publishing, 2018: 801-818.

[15] ZHANG Y, TIAN Y, KONG Y, *et al*. Residual dense network for image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Calgary: IEEE, 2018: 2472-2481.

[16] 杜敏敏, 司马海峰. A-LinkNet: 注意力与空间信息融合的语义分割网络[J]. 液晶与显示, 2022, 37(9): 1199-1208.
DU M M, SIMA H F. A-LinkNet: semantic segmentation net-work based on attention and spatial information fusion[J]. Chinese Journal of Liquid Crystals and Displays, 2022, 37(9): 1199-1208.

[17] 王旭阳, 王常瑞, 张金峰, 等. 基于跨模态交叉注意力网络的多模态情感分析方法[J]. 广西师范大学学报(自然科学版) , 2024，42(2) : 84-93.
WANG X Y, WANG C R, ZHANG J F, *et al*. Multimodal sentiment analysis based on cross-modal cross-attention network [J]. Journal of Guangxi Normal University (Natural Science Edition) , 2024, 42(2) : 84-930.

[18] 孙斌, 江涛, 贾莉, 等. 基于跨模态联合编码的多模态情感分析[J]. 计算机工程与应用. 2024, 60(18): 208-216.
SUN B, JIANG T, JIA L, *et al*. Multimodal sentiment analysis based on cross-modal joint-encoding[J]. Computer Engineering and Applications, 2024, 60(18): 208-216.

[19] CHEN Q, GE T, XU Y, *et al*. Semantic human matting[C] //Proceedings of the 26th ACM International Conference on Multi-media. New York: ACM, 2018: 618-626.

[20] 黄泳嘉, 史再峰, 王仲琦, 等. 基于混合损失函数的改进型U-net肝部医学影像分割方法[J]. 激光与光电子学进展, 2020, 57(22): 74-83.
HUANG Y J, SHI Z F, WANG Z Q, *et al*. Improved U-net based on mixed loss function for liver medical image segmentation[J]. Laser & Optoelectronics Progress, 2020, 57(22): 74-83.

[21] ZHAO H, SHI J, QI X, *et al*. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. Kyoto: IEEE, 2017: 2881-2890.

[22] CHEN L, ZHU Y, PAPANDREOU G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[C] //Proceedings of the European Conference on Computer Vision (ECCV) . Munich: Springer International Publishing, 2018: 801-818.

[23] ZHANG S H, DONG X, LI H, *et al*. PortraitNet: realtime portrait segmentation network for mobile device[J]. Computers & Graphics, 2019, 80: 104-113.

[24] CAO H, WANG Y, CHEN J, *et al*. Swin-unet: unet-like pure transformer for medical image segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[25] 冯广, 潘庭锋, 伍文燕. 基于贝叶斯网络模型的在线学习行为分析[J]. 广东工业大学学报, 2022, 39(3): 41-48.
FENG G, PAN T F, WU W Y. An online learning behavior analysis based on bayesian network model[J]. Journal of Guangdong University of Technology, 2022, 39(3): 41-48.

（责任编辑：杨耀辉　英文审核：熊荣斌）